点击率预估的探索

什么是点击率预估

点击率预测是对每次广告的点击情况做出预测，判断出这次广告会不会被点击，通常以点击的概率的形式给出，就是通常所说的有多大的概率会被点击。点击率预估作为程序化交易必不可少的一部分，一直以来都是优化的重点工作内容。

点击率预估的主要实现

说到预估流程，大体可以分为四个部分

1.特征工程

有这么一句话在业界广泛流传：数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。足可见特征工程的重要性，一直以来点击率预估的核心都在于特征工程。

目前广告点击率预估使用的特征主要由用户特征，广告特征，上下文特征，反馈特征，以及交叉特征这几大部分组成。

1. 用户特征：用户的标签，行为，偏好等等，可以表达一个用户的喜好，这也是个性化推荐的重要特征（好像偏题了，不过广告预估也可以看作一种推荐，只是我们推荐的是用户最可能点的广告）
2. 广告特征：广告的样式，类别等等，这也不用说了，都做广告点击率预估了，没有广告特征实在说不过去
3. 上下文特征：广告的广告位，网络环境，手机平台等环境信息，通常也会对是否会被点击产生很大影响 (反正我在没有wifi的时候是活不下去的，更不用说点广告了)
4. 反馈特征：一个广告过往的点击率情况，可以很直观的反映一个广告的质量
5. 交叉特征：就是之前这些特征的各种各样的组合啦，举一个最简单的例子，两个特征：年龄和性别，可以组合成 年龄_性别 的一个新特征。从理论上而言是为了引入特征之间的交互，也即为了引入非线性，加强了线性模型的表达能力,是有实际意义的。这个特征交叉也是做算法的小哥哥小姐姐们最花脑力和体力的地方。

   想特征是一个脑力加体力的活，更让人郁闷的是，工业界并没有一整套想特征的办法，工业界有的仅仅是验证特征的办法。总的来说选特征的流程，就是先猜想，然后统计验证，然后将特征加到模型中，进行验证。

   对于特征的预处理的方法有很多，具体的实现细节本文也不花费篇幅详解了

2.模型选择和训练

逻辑回归

那么如何预估点击率呢？怎么样去选一个合适的模型？考虑到一个广告只有两种状态，要么被点击，要么不被点击，那么这个预估问题直观上看是一个二分类问题，但是我们给出的预估结果总不能肯定的说这个广告会被点击，或者不会被点击，这样的话这个结果的水分就太大了，我们当然是希望能给出一个概率，那么这里就需要一个回归的模型来解决这个的问题。

说到回归模型，最先想到的当然就是大名鼎鼎的逻辑回归了。而且在传统工业上已经被广泛使用，并且通常都能**取得很好的效果**(此处划重点)。我们当然也不例外(大家都验证好用了，没理由不用呀^^)

逻辑回归 (Logistic Regression, LR),是一种简单的线性算法，y=f(x)，表明自变量x与因变量y的关系，使用sigmoid函数来表达，sigmoid的函数输出是介于（0，1）之间的，符合我们需要的概率的输出

关于逻辑回归的优势和劣势

- 简单，简单，简单，重要的事情说三遍，实现简单，易于理解，实现和解释，计算代价不高，速度很快，存储资源低 ，工业上也很容易实现；单从离线效果上来说，各种gbdt+lr, xgboost甚至是深度学习cnn等等都是碾压单纯的lr的，你说大家为什么实际使用中大多还是选择lr，还不是太复杂了，工程上臣妾实现不了呀TT！当然还有出现bad case的时候，过于复杂的模型只能靠玄学来解释了，这种靠天吃饭的事情，我们还是做不出来的。
- 至于缺点嘛也很明确，就是太简单了，视特征空间内特征之间彼此独立，没有任何交叉或者组合关系，这与实际不符合，比如在预测是否会点击某件t恤是否会点击，如果在夏天可能大部分地区的用户都会点击，但是综合季节比如在秋天，北方城市可能完全不需要，所以这是从数据特征维度不同特征之间才能体现出来的。因此，必须复杂到能够建模非线性关系才能够比较准确地建模复杂的内在关系。一句话说就是线性模型太简单对于非线性的表达能力不够，正所谓成也萧何败也萧何。

3.预估工程

预估工程这里就是通过训练出来的模型，提供出计算点击率的接口，前面都是脑力活，这里就是体现真正技术的时候了，高并发，响应快，怎么厉害怎么来。

实现高性能的核心

1. 优化计算，能算一次的地方绝不算两次，对于一次广告请求有多个广告，再一次请求中用户特征和上下文特征是固定不变的，所以这个算是一个公共部分，可以拆出来单独计算，最后在拼接上广告特征等的结果，可以节约很多的计算资源
2. 缓存，还是缓存，对于计算中的很多反馈特征因为变动的频率并不快，反馈特征按批次更新，目前30s更新一次，而且大量请求中维度有大量相同的地方，所以可以多做缓存，减少对反馈值的请求（这里也是响应时间消耗最多的地方）
3. 优化内存分配，gc等，针对压测中比较慢的方法，消耗资源比较多的函数等等专门优化
4. 其他各种优化方案，是时候展现真正的技术了

4.效果评估

对于模型和特征的效果评估的方法有很多

1. 排序指标，通常会使用auc，auc表达的是模型对排序准确度的评价，这个值越高越好，如auc 80可以简单理解为我们能对80%的数据进行准确排序，
2. logloss物理意义为：衡量预估ctr与实际ctr的拟合程度,越小越好
3. 由于广告排序不单单考虑点击率，还有价格因素在内，所以单单排序准确是不够的，这里还需要值准确，这里就引入一个新的评估指标oe等表达我们预估的值与实际值的比值，这个值越接近1越好，当OE>1时，表示高估，当OE<1时，表示低估
4. 最终指标，收益的提高，所有的一切优化目标都是为了能带来更高的收益，这个当离线各项指标有所提高的时候就可以进行线上的灰度测试了，一般会切一部分流量进行ab-test，只有收益提高了才是真正的提高，其他指标只能作参考

从lr到fm的探索

FM

你们以为这样就完了吗，天真，我的废话还很多呢，哦不是，我们做的事情还多着呢，前面讲的大体流程只是初级阶段，接下来该进化了

前面说到特征工程中最耗费脑力，最耗费体力的就是选择特征交叉了，累死小哥哥小姐姐了。那么有人说全部交叉一遍不就好了，too yong too navie 呀。

特征编码时常用的one-hot编码，会导致特征非常稀疏（很多0值）。常用的特征组合方法是多项式模型

$$y(x) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{ij} x_i x_j$$

其中xi表示第i列特征，n表示特征数，w0,wi,wij为模型参数。模型参数为n^2个。在对模型进行训练时，采用SGD(随即梯度下降)，由于特征较稀疏，大部分wij的梯度值为0，那么参数wij的值就不准确，会影响模型的效果。

上面是官话，翻译过来就是，你们特征太多太多，样本覆盖不够，学习不过来了。还有就是特征爆炸式增长带来的计算量的问题等。

这个时候就需要拿出大杀器了，FM就是为了解决这个问题，FM的基本原理是将这些二项式矩阵做矩阵分解，将高维稀疏的特征向量映射到低维连续向量空间，然后根据内积表示二项式特征关系，复杂度为O(kn^2)

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i \, x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i \, x_j$$

再通过一些公式变换，最终复杂度控制在O(kn)，训练时间复杂度也是O(kn)，也就是线性时间，FM通过对二项式稀疏进行低维连续空间的转换，能够有效地解决多项式中存在的二次项系数在大规模系数数据下不更新的问题，另外由于训练预测复杂度均为线性，这样逻辑下由于要计算多项式和，复杂度是n^2，由于FM的这几个特性，在实际场景中，FM也大规模的应用在CTR中，尤其是在数据极其系数的场景下，FM效果相对于其他算法有很明显提高。

Field-aware FM FMM全称是 Field-aware FactorizationMachine，相对于FM增加了Field信息，每个特征属于一个field，举个例子：用户身份经期，身份辣妈都可以归属在一个Field “身份” 下。其中fi表示特征i所属的field，需要训练的V为n*k*f,f为field的个数

$$y(x) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} < v_{i,f_j}, v_{j,f_i} > x_i x_j$$

由于FFM加入field，使得训练和预测过程参数计算不能简化，复杂度为O(kn^2)

最终效果

在最终实现使用后ffm确实明显较lr有很大提高，在所有发布的版本中提高最多，计算复杂度增加带来的cpu等消耗的曾加和响应时间的增加也在可接受范围内，总体还是很有效的

从cpc到ocpc的探索

顺便扯一下和ctr预估相关的扣费方式，广告投放的收费方式有很多种，常见的有CPM，CPC， CPA等 – CPM cost per Mille 按量付费，我投放多少曝光收多少钱，通常起售价格较高，这是对平台最为有利的方式，通常是较好的位置。投放品牌类广告，追求曝光，对于广告主来说希望的就是更多的曝光，目标是提高品牌效应，品牌知名度等，追求的是长期的效果 – CPC cost per click 按点击付费，通常CPM无法完全消耗流量的情况下，为了不浪费流量，剩下的曝光会再次售卖，这是比较好的方式就是cpc了， cpc通常是追求短期广告效果，对广告主最直观的计量单位就是点击，所以按点击扣费的方式也比较容易被接受，而且按点击扣费投放成本也会更低 – CPA cost per action 按效果

付费，这里的action可以是下载，激活，付费等行为，这也是广告主投放广告的目标，要为广告主带来直接的收益（用户，消费等））

对于大部分的广告主来说投放广告以效果广告居多，按照二八定律，有能力投放品牌广告的毕竟是少说，大多数广告主还是希望能低成本的投放广告，并且能过立竿见影的带来收益，这里最大的诉求当然是cpa的扣费方式，目前市面上比较少看到这种收费模式，对于dsp来说cpa的方式要承担更大的风险，而且对于广告产生的效果也依赖广告主的数据，这个不是很好把控，每增加一道转化就增加一层风险。目前比较有能力做cpa广告的就是淘宝了，因为广告主的销售数据都在马云爸爸手里呀。对于其他的dsp来说可就没那么容易拿到广告主的数据了，拿到了可不可信还不一定呢。

所以广告主希望cpa，dsp平台说我最多做到cpc，这不就得打起来了，金主爸爸不能得罪呀，于是ocpc的投放方式应运而生，ocpc介于cpc和cpa之间，按点击扣费，优化的目标从点击变成广告的转化。

OCPC广告 分为两个阶段 – 第一阶段的主要目的是积累数据，出价公式和普通CPC广告相同。第二阶段按OCPC广告的出价公式参与竞价。把两阶段的规则提前和代理商说明，实际上也就是提前说明了一些投放量比较少，转化数比较少的广告不适合投放OCPC竞价模式。第一阶段，第二阶段的临界点是广告转化数（60个） – OCPC广告第二阶段: 出价公式: cpm = ctr *cvr* cpa出价 ，该阶段的广告效果 在原有cpc（cpm = ctr * cpc出价） 出价公式的基础上多考虑了一个cvr转化率的评估因素,在其他条件对等的情况下优先考虑转化率更高的广告，为广告主带来更好的效果

那么这里就就不单单是要预估ctr了，还有cvr也需要预估，这里可以参照ctr预估的流程，只是这回我们预估的是转化了

总结

这篇文章讲述了广告团队在效果广告优化上的探索，以上内容总结自全体团队成员的共同努力的成果，后续我们还有许多其他的内容需要探索，例如ftrl在线的预估模型等内容，广告反作弊等，希望广告团队以后的内容能越做越好